UNIVERSIDAD DE SALAMANCA

FACULTAD DE PSICOLOGÍA

GRADO EN PSICOLOGÍA

LA ESTADÍSTICA ROBUSTA. NUEVOS CAMINOS EN LA INVESTIGACIÓN

GARCÍA SÁNCHEZ, JENNIFER
Tutor: Delgado Sánchez Mateos, Juan

JULIO DE 2015

Yo, Jennifer García Sánchez, declaro que he redactado el trabajo “La estadística robusta. Nuevos caminos en la investigación”,para la asignatura de Trabajo Fin de Grado en el curso académico 2014/2015 de forma autónoma, con la ayuda de las fuentes bibliográficas citadas en la bibliografía, y que he identificado como tales todas las partes tomadas de las fuentes indicadas, textualmente o conforme a su sentido.

Índice

Índice de tablas

Índice de figuras

RESUMEN

En el presente trabajo se pretende comparar los métodos estadísticos paramétricos y los robustos, con el fin de determinar cuál es más preciso y fiable de cara a los análisis de datos. Para ello nos basaremos en las ideas aportadas anteriormente por diversos autores, por ejemplo Wilcox o Hampel. A lo largo del estudios, se definirán cada uno de los métodos que van a ser empleados, y, posteriormente, se realizará un análisis de una muestra de puntuaciones objeto de estudio, a saber, la tasa de mortalidad infantil de varios países del mundo. Los resultados obtenidos muestran una clara diferencia entre la media y los estadísticos robustos utilizados, lo cual, nos permite concluir que estos últimos se ven menos afectados por las balas perdidas de la distribución, y, por tanto, son más precisos y fiables cuando no se trata de distribuciones normales.

- Palabras clave: estadística paramétrica, estadística robusta, núcleo central de datos, m-estimadores y outliers.

I. INTRODUCCIÓN.

1.1 JUSTIFICACIÓN TEÓRICA.

Para llevar a cabo un análisis preciso de los datos en un estudio, es conveniente conocer cuál es el método más exacto para hacerlo. Por tanto, lo que se pretende a continuación es comparar dos métodos estadísticos, por un lado la estadística paramétrica, la cual, en muchas ocasiones, padece falta de potencia, y por el otro la estadística robusta, con el fin de comprobar cuál es más preciso o más acertado a la hora de hacer un análisis exploratorio o una comparación de datos.

El principal motivo para la realización de este estudio, es que la estadística robusta está en el olvido de la mayoría de los profesionales de este ámbito. Pero sin embargo, teniendo en cuenta que la robustez queda definida como “la propiedad que tiene una prueba estadística cuando sus resultados no son sensibles respecto de las desviaciones de los supuestos básicos de dicha prueba” (Palmer, 1999, p.79), resulta más que interesante conocer cómo funciona y las aplicaciones que tiene, sobre todo cuando se trabaja con muestras que no se adaptan a la normalidad, es decir, la mayoría (por no decir todas) las presentes en el mundo real. Esto se muestra en el artículo de Micceri (1989), donde el autor obtiene como resultado que muy pocas de las 144 distribuciones analizadas parecen ser aproximaciones razonables a la curva normal. Además, recalca la advertencia de Geary (1947)“la normalidad es un mito; nunca hubo, y nunca habrá, una distribución normal”

> “Para empezar, las distribuciones no son nunca normales. […] Creer en la distribución normal implica que solo se necesitan dos números para decirnos todo acerca de las probabilidades asociadas de una variable al azar: la media de la población y la varianza de la población. Lo que es más, asumir normalidad implica que las distribuciones tienen la obligación de ser simétricas” (Wilcox, R., 2005, p.2)

Wilcox añade: “Gauss asume que sí somos capaces de obtener un gran número de observaciones, una gráfica de las observaciones sería simétrica respecto de algún punto desconocido” Al aceptar que las puntuaciones se distribuyen normalmente en torno a un punto concreto, el mismo autor propone que, por tanto, el método más efectivo para estimar este valor es la media.

Sin embargo, no hay ninguna razón para asumir que la media sea el estadístico óptimo, ya que se ha demostrado que hay muchas ocasiones, como el estudio de Laplace en 1818 (extraído de Wilcox, R., 2001), en que la mediana es más precisa. Anteriormente, este mismo autor también encontró algunas situaciones en las que la media no era óptima. De este modo, queda probado que Gauss no tenía razón al asumir que las observaciones siguen una curva normal (Wilcox, R., 2001, p.4)

Tener conocimientos sobre estudios de estadística robusta es muy útil de cara a la exploración y el análisis de datos, ya que, si conocemos los métodos más precisos, nuestras investigaciones y nuestros resultados serán más fiables y acertados.

> “Durante las pasadas décadas se ha empezado a notar, cada vez más frecuentemente, que uno de los más comunes procesos estadísticos (en particular, los optimizados para una distribución normal subyacente) es, por lo visto, excesivamente sensible a las pequeñas desviaciones de los supuestos, y, por tanto, se han propuesto gran cantidad de procesos robustos alternativos” (Huber, 1981, p.1)

Las técnicas estadísticas clásicas de estimación de parámetros, intervalos de confianza y prueba de hipótesis son, en conjunto, denominadas “estadística paramétrica”. Esta asume que la población de la cual la muestra es extraída es *normal* o aproximadamente *normal*, y esta propiedad es necesaria para que la prueba de hipótesis sea válida. Sin embargo, en un gran número de casos no se puede determinar la distribución original ni la distribución de los estadísticos por lo que en realidad no tenemos parámetros a estimar. Tenemos solo distribuciones que comparar. Este es el enfoque conocido como“estadística no paramétrica”.

La principal desventaja de la estadística paramétrica es su falta de potencia en ocasiones en las que existen anomalías en los datos o distribuciones acusadamente no normales. Para evitar este inconveniente podríamos recurrir a la estadística no paramétrica. Pero esta, con *n* muy pequeña, es inconsistente, y con n grande, es aún menos potente que la paramétrica. Por tanto, la alternativa más adecuada veremos que es la estadística robusta.

Para ver más claramente el problema, nos basaremos en las siguientes pruebas usadas, muy frecuentemente, en la estadística paramétrica:

La prueba T de Student o T-test responde a la siguiente fórmula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = MCe = s_e^2 = \frac{SC_e}{p(n-1)}$$

En este caso, cuanto más se disminuye *n*, es decir, el tamaño de la muestra, más fácil es aceptar la H_0. Al ser la *n* menor, la raíz se hace más grande, y por tanto, el denominador también. Esto provoca que el resultado de la división, es decir, de la prueba T, sea más pequeño, y se haga más difícil rechazar la H_0.

La F de Fisher corresponde la siguiente fórmula:

$$F = \frac{s_1^2}{s_2^2}$$

$$\frac{s_1^2}{s_2^2} = \frac{\frac{sc_1}{n_1-1}}{\frac{sc_2}{n_2-1}} = \frac{sc_1}{n_1 - 1} \Big/ \frac{sc_2}{n_2 - 1}$$

Al igual que en el caso anterior, al disminuir*n* resulta más difícilrechazar la H_0, ya que, al ser el divisor de cada cociente más pequeño, el resultado de las divisiones se hace más grande, y el cociente final se reduce.

Y los intervalos de confianza a la fórmula que aparece a continuación:

$$\text{I.C} = \overline{Y} \pm z \times ET_{\overline{y}}$$

$$ET = \frac{s}{\sqrt{n}}$$

En esta fórmula, al disminuir *n* la raíz se hace más pequeña, con lo que disminuye el divisor del cociente, provocando que el resultado de la división sea más grande, es decir, que el error típico (*ET*) sea mayor. Así, multiplicaríamos la *z* por una cantidad más grande y el intervalo de confianza, el cual nos permite rechazar o no la H_0, se hace más amplio y más difícil el rechazo, amén de disminuir la precisión de la estimación al aumentar mucho el error.

En todos estos casos, cuanto más se disminuye *n*, es decir, el tamaño de la muestra, más difícil es rechazar la H_0. Esto indica la falta de potencia de las técnicas paramétricas y la necesidad de buscar otra alternativa. Las técnicas estadísticas no paramétricas ofrecen menor rigidez con respecto a sus condiciones que las técnicas paramétricas, aunque sacrificando para ello su potencia de explicación. Como lo que pretendemos encontrar con técnicas potentes y fiables, nos vemos en la obligación de rechazar estas también. Así, solo nos quedaría la opción de recurrir a las técnicas robustas.

"La estadística robusta, en un sentido amplio, no técnico, tiene que ver con el hecho de que muchas suposiciones comúnmente realizadas en la estadística son, como mucho, aproximaciones a la realidad" (Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W., 1986, p.1). Esto queda latente en los errores de estimación, los "outliers" o"balas perdidas", los problemas de asimetría, las variaciones en la curtosis…El principal problema de las teorías de la estadística paramétrica es que propone procedimientos válidos bajo modelos exactos (normalidad, independencia…), pero no dicen nada acerca de su comportamiento con modelos que se separan de estos supuestos. Los modelos paramétricos - e incluso los no paramétricos - se comportan pobremente incluso bajo pequeñas violaciones de los supuestos asumidos (Hampel, F. et al, 1986, pp.1-2)

> En un sentido amplio, la estadística robusta […] está, en parte, formalizada en “teorías de la robustez”, relativas a las desviaciones de las suposiciones idealizadas en la estadística […] Las teorías de la robustez pueden ser vistas como las teorías de estabilidad de la inferencia estadística” (Hampel, F. et al, 1986, pp. 6-8)

“Una prueba estadística es robusta si se pueden violar sus supuestos sin que ello repercutasubstancialmente en las conclusiones” (SEQC) Para que un estimador sea robusto debe comportarse de manera eficiente y estable sea como sea la muestra de datos con la que se está trabajando.

“Los principales objetivos de la estadística robusta son: describir la estructura que mejor se ajuste al volumen de los datos e identificar puntuaciones desviadas (outliers) o estructuras desviadas para su posterior tratamiento, si se desea” (Hampel, F. et al, 1986, p.11) Una vez que se han determinado las balas perdidas (outliers) de una distribución se puede proceder, mediante otros métodos robustos, al cálculo de los estadísticos de la muestra basándose en métodos que eliminan esas puntuaciones - medias recortadas - o que las asignan otro valor - medias winsorizadas-.

Para datos de alta calidad o, al menos, sin ninguna bala perdida, no es necesario usar métodos robustos. Sin embargo, es difícil asegurar que los datos sean de alta calidad o que hayan sido eliminados correctamente todos los outliers. Algunos investigadores, como por ejemplo Newcomb (1881) citado por Hampel, F. et al (1986) quien decidió hacer un análisis robusto con sus datos y descubrió que ninguno de los valores que obtuvo estaban próximos a la media. Por tanto, aun cuando manejamos datos que creemos que son adecuados para solo usar métodos paramétricos, el empleo de unos buenos métodos robustos para el análisis pueden darnos un incremento notable de la exactitud que obtendríamos usando métodos clásicos (Hampel, F. et al, 1986, pp. 31-32)

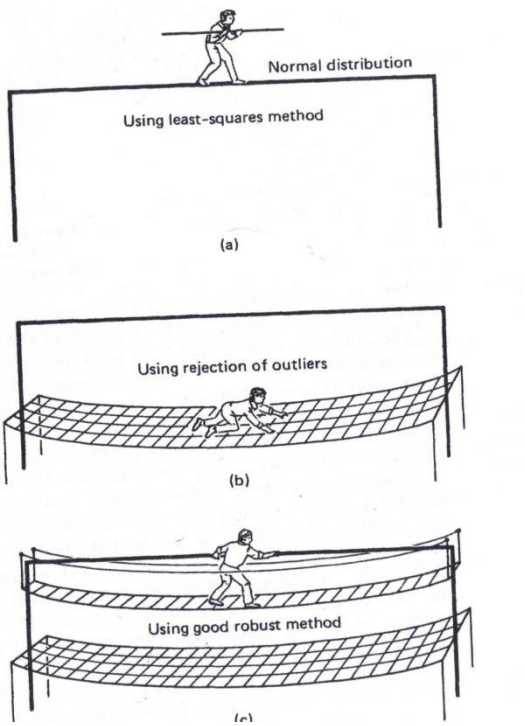

Figura 1: varios modos de analizar datos (tomado de Hample, F. et al, 1986, p.3)

La Figura 1 (Hample, F. et al, 1986, p.3)ilustra muy bien lo que se quiere reflejar en el presente trabajo. En la parte (a), se usa el método de mínimos cuadrados para “atravesar el puente”, este se basa en la optimización matemática, en la precisión, pero deja muy poco margen de error y es muy sensible a las variaciones en la distribución. En la parte (b) se rechazan los valores atípicos, y “el personaje dispone de una red para cruzar al otro lado del puente” hay más margen de error pero se sacrifica la precisión, ya que está muy alejado de la distribución de los datos. En la parte (c), se usan buenos métodos robustos que incluyen la detección de balas perdidas, es el modo más seguro para “que el personaje cruce el puente”, dado que, tiene un “puente más seguro y robusto”, próximo a la distribución y, además, una red que le salvaría si comete algún error. Esta técnica (c) sería la más potente, fiable y precisa a la vez.

Para explicar mejor todo lo dicho anteriormente, pasaremos a relatar un ejemplo práctico con unos datos de Wilcox (2003)

1.2 UN EJEMPLO PRÁCTICO.

En la tabla 1 se muestran las puntuaciones (columna X) que tomaremos como ejemplo para comparar las técnicas clásicas o paramétrica con las técnicas de estadística robusta; la distancia de cada puntuación a la mediana, en valor absoluto (columna |X-M|); la distancia de cada puntuación a la media dividida entre MADN (columna [X-M]/MADN): las puntuaciones que quedarían excluidas, por considerarse outliers, siguiendo el criterio K>|1,28| (columna k>|1,28|); y las puntuaciones que se tienen en cuenta para calcular la media winsorizada al 20% (columna puntuaciones para media winsorizada)

Tabla 1: Puntuaciones Wilcox

X	|X - M|	(X - M)/MADN	k > |1,28|	Puntuaciones para media winsorizada
77	185	-1,09		88
87	175	-1,04		88
88	174	-1,03		88
114	148	-0,88		114
151	111	-0,66		151
210	52	-0,31		210
219	43	-0,25		219
246	16	-0,09		246
253	9	-0,05		253
262	0	0,00		262
296	34	0,20		296
299	37	0,22		299
306	44	0,26		306
376	114	0,67		376
428	166	0,98		428
515	253	1,50	**	515
666	404	2,39	**	666
1310	1048	6,20	**	666
2611	2349	13,90	**	666

Antes de pasar a explicar los estadísticos que usaremos, es necesario explicar qué entendemos como “núcleo central de datos”. Éste estaría formado por los valores de la distribución que se sitúan próximos entre sí dando lugar al “corazón” de las puntuaciones. En la figura 2 se muestran dos formas de representar los datos de una distribución bivariada, una gráfica de contorno (“contour”) y otra de contorno coloreado (“filledcontour”). En ellas se aprecia perfectamente el núcleo central de datos y las puntuaciones alejadas del mismo.

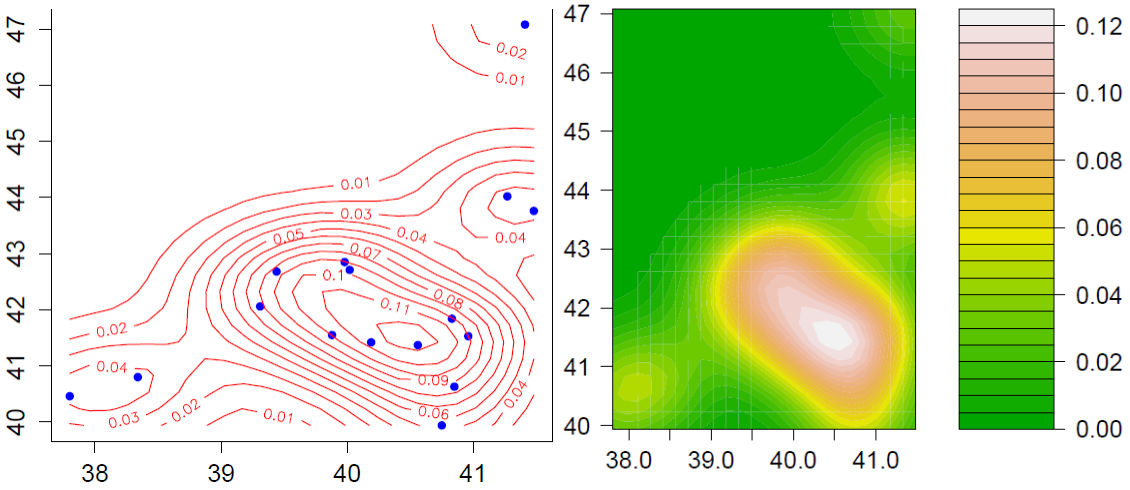

Figura 2: representación de los datos de una distribución.

A partir de los datos anteriormente presentados, calcularemos varios estadísticos:

- Media: la media es aquel valor de la variable que representa el centro de gravedad de la distribución. Su valor se obtiene sumando los valores de la variable y dividiendo por el número de valores totales de dicha variable (Palmer, 1999, p.22) Responde a la fórmula:

$$m = \frac{\sum X_i}{n}$$

- Mediana: es el valor de la variable que divide a la distribución en dos partes iguales conteniendo cada una de ellas el 50% de las observaciones (Palmer, 1999, p.25)

- Media recortada: la media recortada queda definida por la proporción de casos, α, que son excluidos desde cada extremo de la muestra ordenada. Una vez que se han eliminado los valores indicados de cada extremo, se calcula el promedio de los valores restantes. Si α es un múltiplo de 1/n, se eliminan de cada extremo un número entero de valores [αn] y la media recortada es el promedio de los valores restantes, que se calcula mediante la fórmula:

$$m = \frac{\sum X_i}{n}$$

Si α no es múltiplo de 1/n, se elimina un número entero de valores [αn] de cada extremo, y al mayor y menor valor restante se le pondera mediante el peso:

$$p = 1 + [\alpha n] - \alpha n$$

Para calcular la media ponderada, en este caso, se usa la fórmula:

$$T(\alpha) = pXi + Xi+1 + \ldots + Xs-1 + pXs / n(1-2\alpha)$$

(Palmer, 1999, pp.85-86)

Esto es así porque para calcular la media recortada se eliminan las puntuaciones más alejadas del núcleo central de la distribución por ambos lados, y se trabaja solo con el centro de las puntuaciones, así la media no se ve afectada por las puntuaciones extremas.

- Media winsorizada: esta media sustituye los casos excluidos del análisis por el último valor, en cada extremo, que si forme parte del análisis. Cuando se cambian estos valores, se calcula el promedio de las puntuaciones (Palmer, 1999, p. 86) Se calculan usando la fórmula:

$$W(\alpha) = \frac{\sum X_i}{n}$$

donde α es la proporción de casos excluidos por cada extremo.

En este caso no se recortan las puntuaciones extremas, como en el caso anterior, sino que, para mantener el mismo tamaño muestral, lo que se hace es sustituir las puntuaciones que se encuentran fuera del núcleo central de la distribución por otras que si están incluidas en él (la más próxima a la puntuación o puntuaciones eliminadas)

- Desviación típica: la desviación típica calcula un promedio de diferencias de las puntuaciones con respecto a su media (Carro, J. 1994, p.74) Para obtener la desviación típica se usa la fórmula:

$$S_x = \frac{\sum(x - \bar{x})^2}{\sqrt{n}}$$

- Desviación típica Winsorizada: es lo mismo que la desviación típica, descrita anteriormente, pero para su cálculo nos basamos en la media winsorizada en lugar de en la aritmética. La D.T Winsorizada es la raíz cuadrada de la varianza winsorizada, que puede obtenerse por medio de:

$$S_w^2 = \frac{\sum[x_i - W(\alpha)]^2}{n - 1}$$

$$S_w = \sqrt{S_w^2}$$

- MAD: es la mediana de las desviaciones absolutas respecto de la mediana. Para obtener este valor, se calcula X - mediana |X|, la mediana de estas diferencias es MAD.

$$\text{MAD} = \text{Mediana de } |X_i - \text{M}|$$

- MADN: es el MAD normalizado, se usa para estimar σ_{ps} (pseudo desviación estándar) Cuando usamos el programa estadístico R, MAD = MADN.

$$\text{MADN} = \frac{MAD}{0{,}6745}$$

- M-estimador de un paso: un M-estimador se define como Maximun Likelihood Estimator (estimador de máxima verosimilitud) Su objetivo es buscar un índice de localización a partir del conjunto de observaciones, ponderando a éstas en función de lo cerca o lejos que se encuentren del centro de datos (Palmer, 1999, p. 122) Para el cálculo del M-estimador de una muestra se corta la cola con puntuaciones anormalmente distantes del centro de la misma. Hay muchos tipos de M-estimadores, el que usaremos en este ejemplo será el de Huber. Este autor usa el valor de constante K = 1,28 (que corresponde aproximadamente a la *z* para el 80% central de los datos, un núcleo central con un recorte del 20%) para determinar las balas perdidas de la muestra. Por tanto, cualquier valor será declarada un outlier si:

$$|X - M|/MAD/0{,}6745 < 1{,}28$$

Hay varias clases de estimadores (L, M y R), pero los M-estimadores son algunos de los más utilizados y estudiados, ya que, permiten eliminar las balas perdidas siguiendo un criterio basado en la distancia desde cada observación al centro de los datos.

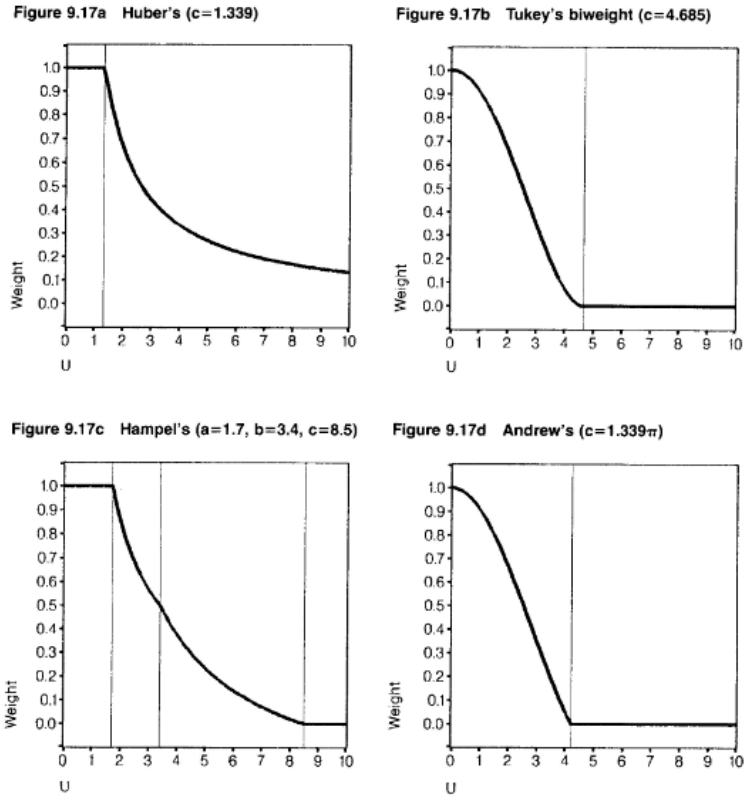

Figura 3: M-estimadores más usados

En la figura 3 se presentan los M-estimadores más usados (Huber, Tukey o bicuadrado, Hampel y Andrew) Norušis (1990) Las líneas verticales de cada gráfica indican los puntos de corte de las distribuciones y son los valores que se tiene en cuentan para calcularlos. El M-estimador de Huber es el único asintótico respecto del eje X (nunca toma el valor y=0), es decir, nunca elimina puntuaciones muy lejanas.

Después de definir los estadísticos que vamos a tener en cuenta a la hora de analizar los datos de Wilcox (Tabla 1), podemos comenzar con el estudio.

Se presentan los resultados (en la Tabla 2), y según estos, podemos afirmar que la media es el estadístico que más se ve afectado por las puntuaciones extremas (515, 666, 1310 y 2611) ya que, estas se sitúan en la parte superior de la distribución y la media tiende a acercase también a este polo:

Tabla 2: Resultados del análisis de los datos de Wilcox.

Estadístico	**Valor**
Media	448,11
Mediana	262
Media recortada al 20 %	342,71
Media winsorizada al 20 %	312,47
Estimador desviación típica población	594,63
Desviación típica Winsorizada	194,39
MAD	114
MADN	169,01
M-estimador de un paso (basado en Huber)	285,16

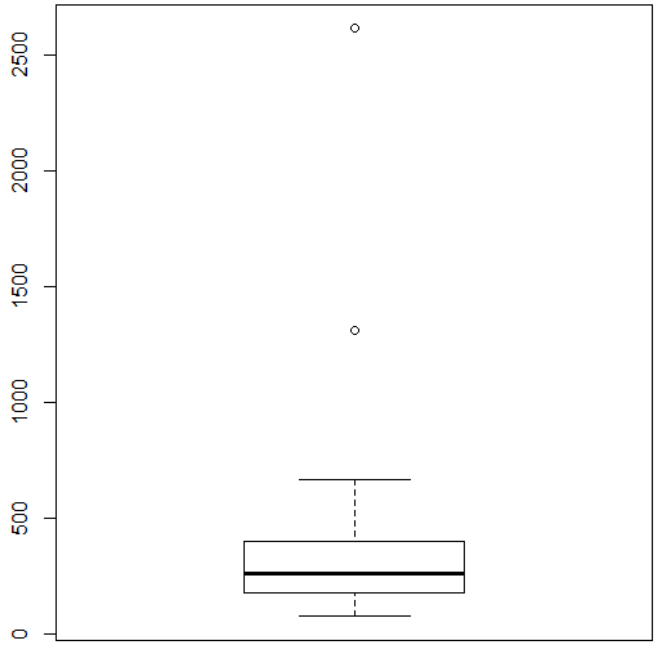

Figura 4: Gráfico de caja y patillas de los datos de Wilcox

En la Figura 4 se podemos ver el gráfico de caja y patillas de la distribución. La caja representa el “núcleo central de datos”, este rectángulo está dividido por un segmento vertical que indica donde se posiciona la mediana. Las puntuaciones que no se encuentren dentro del rango delimitado por las patillas son consideradas “outliers”. En nuestro caso tenemos 2 puntuaciones que se sitúan por encima de la patilla superior.

Antes se ha dicho que la media se ve afectada por cuatro puntuaciones extremas, que serían los outliers siguiendo el criterio de Huber para la detección de balas perdidas:

$$|X - M|/MAD/0{,}6745 < 1{,}28$$

Esto puede verse en la columna 4º de la Tabla 1 (página 7), los 4 valores que tiene asteriscos (**) en esta columna son los que se han identificado como outliers. Sin embargo, si nos fijamos en el gráfico de caja y patillas (Figura 2), nos encontramos solo con 2 balas perdidas: 1310 y 2611. Dependiendo de la robustez y la potencia del método que escojamos a la hora de comprobar que valores son outliers, obtendremos unos u otros resultados.

En la figura 3 se presenta un histograma con la distribución de los datos. Más que para observar cómo se sitúan los datos, donde se puede ver claramente que la puntuación 2611 estaría muy alejada del groso de los valores, lo que se trata de plasmar son las diferencias entre los valores de los estadísticos robustos utilizados en el ejemplo (mediana, media winsorizada al 20%, media recortada al 20% y m-estimador) y el estadístico de tendencia central, la media.

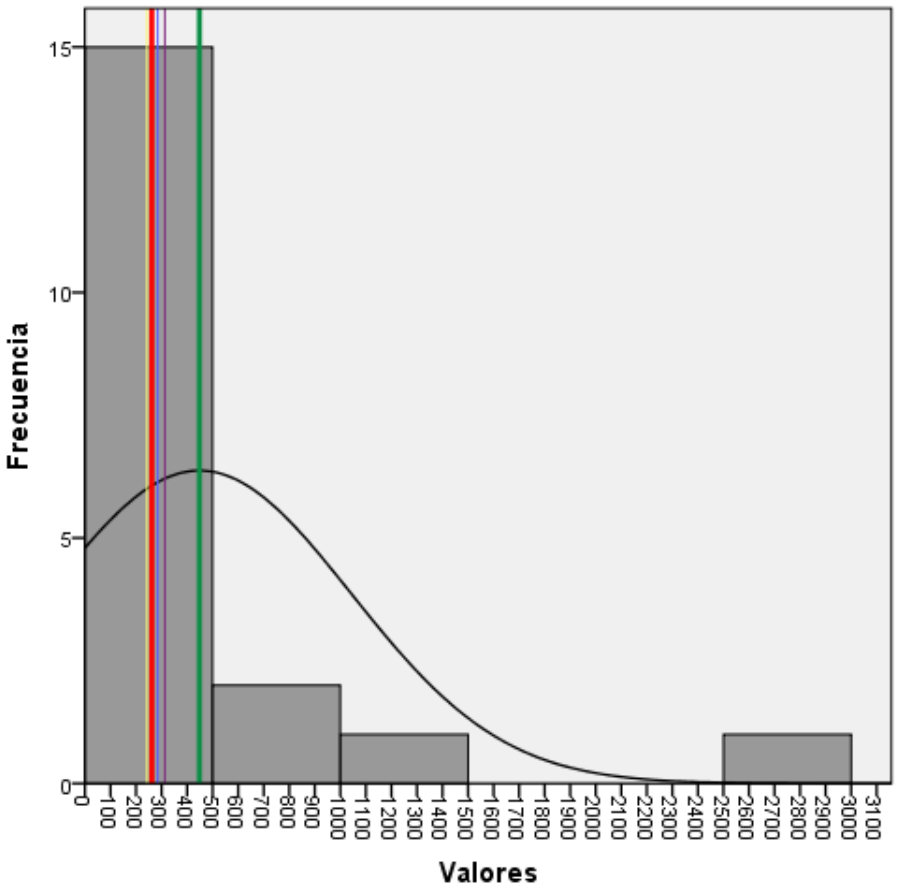

Figura 5: Histograma de los datos de Wilcox

La línea verde representa la media de las puntuaciones (448,11), la morada la media recortada al 20% (342,71), la azul la media winsorizada al 20% (312,47), la roja la mediana (262) y la amarilla el m-estimador basado en Huber (285,16)

Como se puede apreciar, todos estos valores están bastante próximos, excepto la media, que se distancia más del resto de estadísticos debido a la influencia de las puntuaciones extremas.

El estimador de la desviación típica de la población (594,63) y la desviación típica Winsorizada (194,39), también son tan diferentes debido a la forma por la cual se los calcula. En el primer caso se tiene en cuenta la media y en el segundo la media winsorizada, al estar esta última menos afectada por los outliers se obtiene un valor más pequeño, más robusto y potente.

1.3 OBJETIVOS.

El objetivo fundamental de este estudio es mostrar dos métodos de análisis de datos, el clásico o paramétrico y el robusto, y hacer una comparación entre ellos con el fin de comprobar cual se comporta mejor en situaciones en las que los datos no se distribuyen de manera normal (que suele ser siempre)

Si se demuestra que el método robusto es mejor, se estará impulsando una vía diferente de la que se suele usar en las investigaciones hoy en día, que llevaría a unos resultados más precisos, potentes y fiables.

II. METODOLOGÍA.

2.1 DATOS.

Para el estudio usaremos como datos la tasa de mortalidad infantil por cada 1000 nacidos vivos, tomada en 2013, de una serie de países del mundo, a saber: Alemania, Angola, Argentina, Australia, Bahamas, Brasil, Canadá, China, Colombia, Costa Rica, Cuba, Egipto, El Salvador, España, Estados Unidos, Francia, Grecia, Hungría, Israel, Japón, Kuwait, México, Nueva Zelanda, Palau, República Centroafricana, Samoa, Sierra Leona, Tailandia, Tonga y Túnez.

Estos datos han sido tomados de la página web del Grupo Banco Mundial (http://datos.bancomundial.org/indicador/SP.DYN.IMRT.IN), una organización que busca acabar con la pobreza extrema y promover la prosperidad compartida.

Se decidió tomar 5 países de cada continente, pero como en la Antártida no habitan personas, se dividió en dos partes el continente americano: América Norte-Central y América del Sur. Para ello, se separaron los países por grupos y llevo a cabo la selección. La muestra de datos se ha tomado basándose en el método de muestreo aleatorio simple o sin reposición, cada país tenía la misma probabilidad de ser elegido, y, cuando un país resultaba elegido no se tenía en cuenta para la siguiente extracción.

Al final se obtuvo una muestra de 30 países con sus correspondientes datos de mortalidad infantil por cada 1000 nacidos (x[media] = 18’03, d.t = 28’76, mínimo = 2, máximo = 107)

2.2 MATERIALES.

Usaremos los datos definidos anteriormente, extraídos de la página web del Grupo Banco Mundial, que se tratarán con dos programas estadísticos diferentes: IBM SPSS Statistics 20 y R (Rstudio y Rcommander); y con el programa Microsoft Excel Starter 2010.

2.3 PROCEDIMIENTO.

El desarrollo del trabajo se llevará a cabo de la siguiente manera:

1) Se introducirán los datos en una nueva hoja de cálculo Excel y se realizará la tabla con los análisis pertinentes.

2) Se introducirán los datos en una nueva hoja de datos SPSS y se llevarán a cabo los análisis que se han fijado (IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.)

3) Se introducirán las puntuaciones en el programa RStudio y se realizarán los análisis oportunos (R Development Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-proyect.org.)

Los análisis que se van a realizar están definidos en el próximo apartado del presente trabajo (2.4 Análisis estadísticos). Una vez realizados todos ellos, se seleccionarán y se presentarán de la forma más adecuada posible.

Para el análisis de los datos se usarán los siguientes estadísticos:

- ✓ Media
- ✓ Desviación típica
- ✓ Mediana
- ✓ Trimedia
- ✓ Media recortada al 20%
- ✓ Media winsorizada al 20%
- ✓ Desviación típica winsorizada
- ✓ M-estimador de Huber
- ✓ M-estimador de un paso modificado
- ✓ M-estimador de Tukey
- ✓ M-estimador de Hampel
- ✓ M-estimador de Andrews
- ✓ M-estimador de Pitman

Para la detección de los posibles “outliers” se usará un diagrama de tallo y hojas, un gráfico de caja y patillas (“boxplot”), una modificación del mismo (“outplot”) y la ecuación ya propuesta anteriormente:

$$|X - M|/MAD/0{,}6745 < 1{,}28$$

También se realizará un histograma para comprobar de manera más visual la distribución de las puntuaciones e indicar donde se encuentra cada uno de los estadísticos calculados, a fin de compararlos.

2.4 ANÁLISIS ESTADÍSTICOS.

Con el programa SPSS se calcularan los siguientes estadísticos: M-estimador de Tukey, M-estimador de Hampel y M-estimador de Andrews. Además, en un análisis exploratorio de los datos, se muestran algunos estadísticos interesantes para el estudio.

Con Excel se elaborará la tabla de presentación de los datos y, además, la distancia, en valor absoluto, de cada puntuación a la mediana y las puntuaciones que se tendrán en cuenta para calcular la media winsorizada. A parte de esto, se presentará una columna en la que queden identificados los outliers basándose en el método de detección de balas perdidas propuesto por Huber. Se calcularán también algunos estadísticos de interés.

Con R, RStudio y RComander se realizarán el resto de análisis y gráficos.

.

III. RESULTADOS Y DISCUSIÓN.

3.1 RESULTADOS.

En la tabla 3 se muestran los datos que van a ser sometidos a los análisis estadísticos. Además, se ha calculado la distancia a la mediana de cada puntuación (en valor absoluto) para hallar MAD, que tiene un valor de 5, y posteriormente MADN, que tiene un valor de 7,41. A partir de estos datos, podemos definir que valores son considerados outliers mediante la fórmula:

$$|X - M|/MAD/0{,}6745 < 1{,}28$$

Estos han sido marcados con dos asteriscos en la 5ª columna de la tabla. También se presentan las puntuaciones que se van a usar para el cálculo de la media winsorizada.

Tabla 3: Puntuaciones tasa de mortalidad.

País	Tasa de mortalidad infantil por 1000 nacimientos vivos	|X-M|	(X-M)/MADN	k > |1,28|	Puntuaciones para media winsorizada
Alemania	3	7	-0,9443		3
España	4	6	-0,8094		3
Francia	4	6	-0,8094		3
Grecia	4	6	-0,8094		3
Hungría	5	5	-0,6745		4
Japón	2	8	-1,0792		4
Tailandia	11	1	0,1349		4
China	11	1	0,1349		5
Israel	3	7	-0,9443		5
Kuwait	8	2	-0,2698		5
Argentina	12	2	-0,2698		5
Bahamas	10	0	0		6
Brasil	12	2	-0,2698		8
Colombia	15	5	-0,6745		8
Cuba	5	5	-0,6745		10
Canadá	5	5	-0,6745		10
EEUU	6	4	-0,5396		11
México	13	3	0,4047		11
El Salvador	14	4	-0,5396		12
Costa Rica	8	2	-0,2698		12
Australia	3	7	-0,9443		13
Nueva Zelanda	5	5	-0,6745		13
Palau	15	5	-0,6745		14
Samoa	16	6	-0,8094		15
Tonga	10	0	0		15
Sierra Leona	107	97	13,0853	**	16
Angola	102	92	12,4108	**	16
Rep Centroafricana	96	86	11,6014	**	16
Egipto	19	9	1,2141		16
Túnez	13	3	0,4047		16

Tabla 4: Estadísticos de la muestra.

			Estadístico	Error típ.
TasaMortalidad	Media		18,03	5,251
	Intervalo de confianza para la media al 95%	Límite inferior	7,29	
		Límite superior	28,77	
	Media recortada al 5%		14,06	
	Mediana		10,00	
	Varianza		827,137	
	Desv. típ.		28,760	
	Mínimo		2	
	Máximo		107	
	Rango		105	
	Amplitud intercuartil		10	
	Asimetría		2,690	,427
	Curtosis		5,952	,833

Con los estadísticos que se muestran en la tabla 4 ya nos podemos hacer una idea de cómo se distribuyen nuestros datos. La media (18,03) se encuentra bastante alejada de la mediana (10) y de la media recortada al 5% (14,06 - ¡ojo! Al 5%, luego la recortaremos más, pero el mero hecho de eliminar un valor alejado por cada lado ya hace variar la media en 4 puntos) La desviación típica de la muestra es muy grande (28,76), de hecho, al restar una d.t a la media ya nos vamos a valores negativos. Los índices de asimetría y curtosis nos indican que estamos ante una muestra poco normal, para que lo fuera sus valores deberían ser 0 y 3 respectivamente.

Tabla 5: M-estimadores de las puntuaciones.

	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
TasaMortalidad	9,40	8,60	8,67	8,60

a. La constante de ponderación es 1,339.

b. La constante de ponderación es 4,685.

c. Las constantes de ponderación son 1,700, 3,400 y 8,500.

d. La constante de ponderación es 1,340*pi.

Todos los M-estimadores calculados (véase tabla 5) tienen un valor parecido, de hecho el de Tukey y el de Andrews tienen el mismo valor. El de Huber se separa un poco del resto de valores, pero se sitúan entre estos y la mediana de la distribución.

A continuación calculamos con el software R algunos estadísticos convencionales. En primer lugar, introducimos los datos en R:

```
>x <- c(3, 4, 4, 4, 5, 2, 11, 11, 3, 8, 12, 10, 12, 15, 5, 5, 6, 13, 1
4, 8, 3, 5, 15, 16, 10, 107, 102, 96, 19, 13)
```

Calculamos después su media:

```
  >mean (x)
[1] 18.03333
```

El intervalo de confianza al 95%:

```
     >conf.level=0.95
95 percent confidence interval:
7.294178 28.772489
```

después su desviación típica:

```
  >sd (x)
[1] 28.75999
```

finalmente, el error típico (o estándar) de la media:

```
>sd(x)/sqrt(30)
[1] 5.250831
```

Pedimos también un resumen general de los datos:

```
             >summary (x)
Min. 1st Qu.Median    Mean 3rd Qu.    Max.
2.00    5.00   10.00   18.03   13.75  107.00
```

que nos permite calcular la trimedia de las puntuaciones:

```
>5/4 + 10/2 + 13.75/4
   [1] 9.6875
```

Si comparemos los resultados obtenidos con la media, su error típico, calculado con la desviación típica y los resultados que proponemos a continuación, veremos alguna de las ventajas del uso de la estadística robusta:

Antes de nada es conveniente saber que para calcuar los siguientes estadísticos es ne cearios cargar algunos paquetes adicionales en R: Rcmdr, splines, RcmdrMisc, car, sand wich, robust, fit.models, MASS, robustbase, rrcov, Smoothmest, WRS y akima.

En el siguiente bloque se muestra el valor de la mediana de la distribución, sus inter valos de confianza y su error estándar.

```
>median (x)
[1] 10

>msmedci (x)
$ci.low
[1] 6.575923
$ci.hi
[1] 13.42408

>msmedse (x)
[1] 1.74701
```

Mientras el intervalo de confianza alrededor de la media de encuentra entre los valor es7.29 y 28.77, excesivamente amplio, el intervalo de confianza alrededor de la mediana es mucho más preciso (entre 6.58 y 13.42). Por otra parte, el error típico de la media es igual a 5.25 mientras que el de la mediana es 1.75, aproximadamente un tercio del prime ro.

A continuación, se presenta la media recortada al 20% (observese que ha cambiado bastante el valor respecto a la media recortada al 5% presentada anteriormente), susinter valos de confianza y su error estándar.

```
>mean(x,trim=20/100)
[1] 9.277778

>trimci(x)
[1]  6.399349 12.156206
$estimate
[1] 9.277778

>trimse (x,tr=.2)
[1] 1.364303
```

Los valores de la media recortada, de la mediana y de la media son respectivamente iguales a 9.28, 10 y 18.03. Obviamente, solo la media se ve afectada por las puntuacione s anómalas, y los estimadores robustos son más parecidos, aunque cada uno de ellos se vasa en diferentes supuestos, como hemos puesto de manifiesto más arriba cuando se de finieron.

A continuación se indica el valor de la media winsorizada al 20% (misma proporció n que para la media recortada), su varianza y desviación típica, su intervalo de confianza y su error estándar.

```
>winmean (x,tr=.2)
[1] 9.366667

>winvar (x,tr=.2)
[1] 20.1023

>sqrt (20.1023)
[1] 4.483559

>winci(x)
[1]  6.42051 12.31282

>winse(x)
[1] 1.396405
```

Nuevamente, el estimador central (en este caso, la media winsorizada) arroja un valo r coherente con los anteriormente obtenidos y separado del estimador “media aritmética ”. Del mismo modo, tanto el intervalo de confianza como el error típico son coherentes con los otros valores robustos alculados.

A partir de aquí se muestran los M-estimadores calculados. El M-estimador de un pa so (onestep) y el de un paso modificado con su intervalo de confianza correspondiente. El M-estimador de Huber (mest) que es el mismo que el de un paso, su intervalor de con fianza y su error estándar. Y el M-estimador de Pitman. Nótese que el valor para el M-e stimador de Huber en este caso - 9,795 - difiere del obtenido con el programa SPSS - 9, 40 - dado que usan otra constante para su cálculo.

```
>onestep (x)
[1] 9.795034

>mom (x)
[1] 8.740741

>momci(x)
[1]  5.50000 10.73077

>mest (x)
[1] 9.795034

>mestci (x)
[1]  6.800539 12.690560

>mestse (x)
[1] 1.236079

>pitman (x)
[1] 9.569955
```

Hasta aquí hemos visto dos tipos de estadísticos. Por un lado la media aritmética, co n su error típico y sus intervalos de confianza, la desviación típica y la varianza, que se vería gravemente afectada por las puntuaciones extremas. Y por otro lado los métodos r obustos: mediana (su error típico y su intervalo de confianza), media recortada (su error típico y su intervalo de confianza, media winsorizada (su error típico, su intervalo de co nfianza y la varianza y la desviación típica winsorizada) y m-estimadores (Huber [su err or típico y su intervalo de confianza] , Hampel, Andrew, Tukey y pitman). Estos se refie ren al “núcleo central de datos”, no consierando las puntuaciones extremas (media recor tada) y otras optimizando la influencia de estas (media winsorizada)

3.1.1 “Outliers”

A continuación se presenta uno de los métodos elegidos para detectar las balas perdi das de la distribución, el procedimiento “outbox” de Wilcox (Wilcox, R., 2003, p.81). A diferencia del gráfico de caja y bigotes, o “boxplot”, “outbox” utiliza para su cálculo la mediana de la distribución. Así se ataja el problema por el cual los “boxplot” eran critic ados, que la proporción de numero que son declarados outliers depende del tamaño de la muestra (Wilcox R., 2003, p.80) El apartado $out.val nos indica las balas perdidas y $ke ep los valores que forman el núcleo central de datos.

```
>outbox (x)
  $out.val
[1] 107 102  96

  $keep
[1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
23 24 25 29 30

  $n
[1] 30

$n.out
[1] 3
```

En la figura 6 se puede observar la gráfica de caja de la distribución de las puntuaciones. La distribución sería prácticamente normal de no ser por las tres balas perdidas: Sierra Leona, Angola y República Centroafricana, que se sitúan bastante por encima de los bigotes del gráfico.

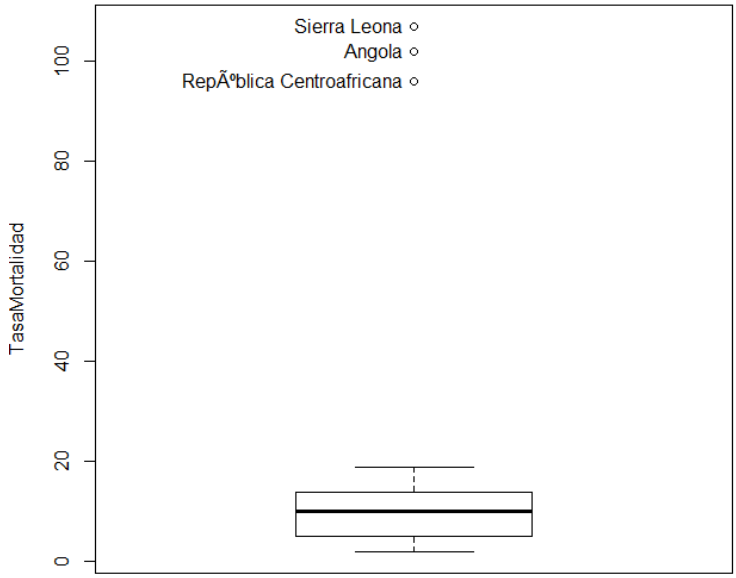

Figura 6: gráfico de caja y bigotes de la distribución.

En la gráfica de comparación de cuartiles (Figura 7) se puede apreciar como todas las puntuaciones se encuentran entre las dos líneas rojas discontinuas (datos normales) menos las de Sierra Leona, Angola y República Centroafricana, que serían los “outliers”.

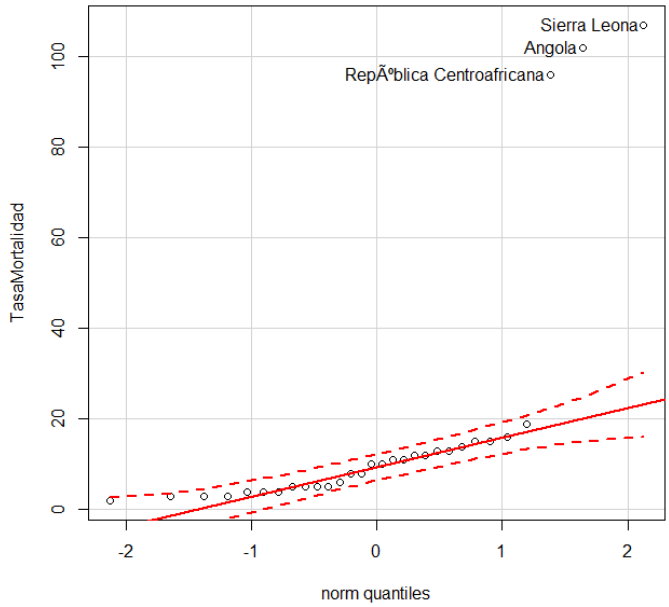

Figura 7: Gráfica de comparación de cuartiles de la distribución

En el gráfico de tallo y hojas (véase figura 8) tambien observamos tres balas perdidas, los valores 96, 102 y 107 que corresponde a Sierra Leona, Angola y República Centroafricana. La “hoja” que se sitúa entre parentesis nos está indicando el lugar de la mediana (10)

```
1 | 2: represents 12
 leaf unit: 1
            n: 30
   4      t | 2333
  11      f | 4445555
  12      s | 6
  14     0. | 88
  (4)    1* | 0011
  12      t | 2233
   8      f | 455
   5      s | 6
   4     1. | 9
HI: 96 102 107
```

Figura 8: Gráfico de tallo y hojas de la distribución.

Mediante los cinco métodos que hemos empleado para detectar las balas perdidas: ecuación de Huber, “outbox”, “boxplot”, comparación de cuartiles y tallo y hojas; hemos obtenido los mismos outliers, los valores correspondientes a los países: Sierra Leona, Angola y República Centro africana.

Por último, en el gráfico de la densidad de los datos (figura 9) se pueden diferenciar claramente dos grupos de datos: el núcleo central de la distribución (parte izquierda) y las balas perdidas (parte derecha) En ese gráfico se indica, mediante líneas verticales, donde se sitúan los principales estadísticos calculados. La media es la línea que se encuentra más separada de las otras líneas, las cuales representan la mediana, el m-estimador de Huber, la media recortada, la trimedia y la media winsorizada.

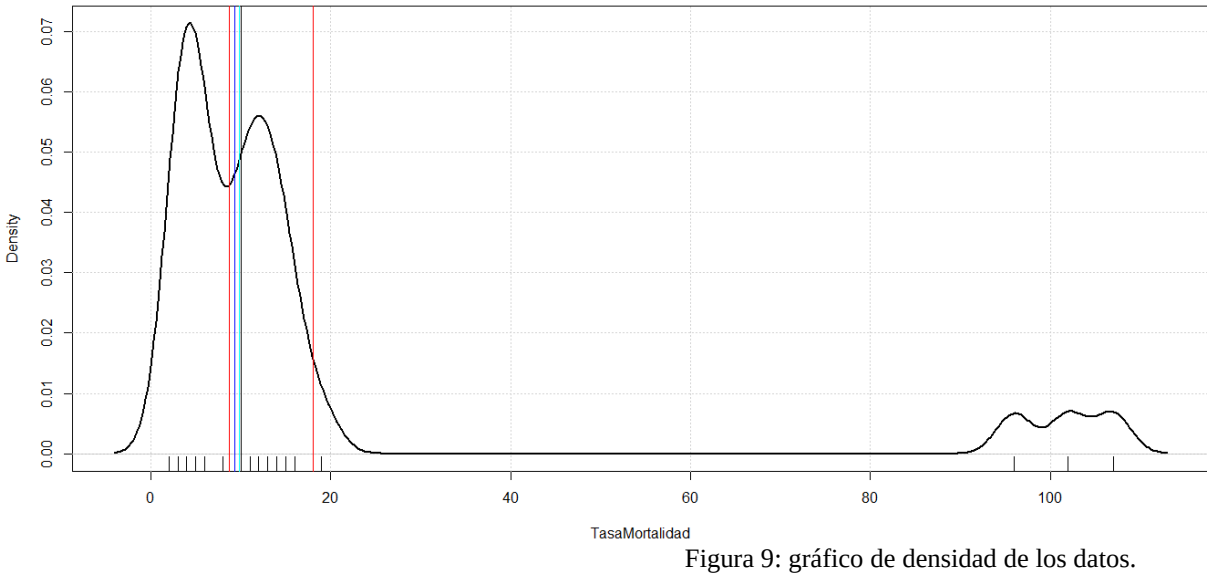

Figura 9: gráfico de densidad de los datos.

3.2 DISCUSIÓN.

Como se relata al comienzo del presente trabajo, de lo que trata es de comparar dos métodos estadísticos, el paramétrico y el robusto. A lo largo del marco teórico, se presenta como “mejor”, es decir, más preciso y fiable, el método robusto, como así ha resultado ser en nuestros resultados.

La Figura 9 podría considerarse un resumen de los análisis estadísticos llevados a cabo, y en ella, podemos ver claramente como los resultados obtenidos los métodos robustos se ven menos afectados por la incidencia de las puntuaciones que se separan del núcleo de datos.

Tal y como sucedió en el estudio de Laplace, anteriormente descrito, la mediana resulta ser más precisa que la media, y, este es un caso la media no es precisa debido a las puntuaciones extremas. Con esto, también podemos rebatir la idea de Gauss que le llevó a asumir que las observaciones seguían la curva normal, ya que, en este caso, por ejemplo, no es así. Como vemos también en la Figura 9, la distribución de las puntuaciones se aleja mucho de la distribución normal

Según los resultados de nuestro estudio tenemos tres balas perdidas, las puntuaciones correspondientes a los países: Sierra Leona, Angola y República Centro africana. Los estadísticos robustos calculados se ven poco afectados por estas puntuaciones extremas, ya que, se calculan teniendo en cuenta el núcleo central de datos, no como la media que toma todas las puntuaciones de las distribución, por ello, tienen un valor más alto que los demás estadísticos, porque los valores altos tiran de ella hacia valores más altos. Cuando más puntuaciones extremas recortemos, es decir, no tengamos en cuenta, menos se vera afectada la media. Una prueba de ello es la media recortada al 5%, cuyo valor es 14,06 y la media recortada al 20%, cuyo valor es bastante menor, concretamente 9,277778.

En nuestro estudio sucede como en el de Newcomb citado anteriormente, al hacer un análisis robusto de los datos se descubrió que ninguno de los valores que obtenidos estaban próximos a la media. Esto nos lleva, al igual que en su caso, a un incremento de la exactitud que obtenemos con los métodos paramétricos.

En lo referente a los intervalos de confianza y al error estándar de cada estadístico calculado, podemos observar que los correspondientes a los estadísticos robustos se sitúan cercanos a 6 y 12, aproximadamente, en el caso de los I.C y entre 1,2 y 1,8 en el caso del error estándar. Sin embargo, en el caso de la media, el I.C corresponde a los valores 7,3 y 28,9 y el error estándar es 5,25. Los de los estadísticos robustos son más precisos y aproximados que los de los estadísticos paramétricos, la media, ya que tienen en cuenta, únicamente, el núcleo central de datos.

IV. CONCLUSIONES Y PROSPECTIVA.

4.1 CONCLUSIONES ESPECÍFICAS.

Si nos basamos en el análisis de los resultados obtenidos, podemos realizar varias conclusiones:

- La media es menos precisa que los estadísticos robustos.
- Es imprescindible abordar un análisis exploratorio de datos antes de intentar ningún análisis estadístico paramétrico.
- Es necesario comprobar si una distribución de puntuaciones tiene puntuaciones anómalas, huecos en la distribución, anomalías en la distribución, etc. con el fin de elegir los mejores métodos para su análisis.
- La estadística robusta proporciona estimadores que se ven menos afectados por las puntuaciones extremas, proporcionan intervalos de confianza más ajustados y errores estándar inferiores.
- El M-estimador de Huber nos proporciona una puntuación más pareja al resto de los estadísticos robustos calculados. Es el M-estimador más adecuado y el que representaría adecuadamente a los demás.

4.2 CONCLUSIÓN GENERAL.

En el presente trabajo hemos mostrado dos métodos de análisis de datos, el paramétrico, mediante el uso de la media, y el robusto, con el resto de estadísticos. En la comparación de los mismos hemos podido comprobar que los métodos robustos se comportan mejor cuando nos encontramos con datos que no se distribuyen normalmente, como los de nuestro estudio.

Por tanto, teniendo en cuenta los resultados obtenidos, podemos concluir que la estadística robusta ofrece mayor exactitud y precisión que la paramétrica, dado que la 1ª trabaja con el núcleo central de datos, sin tener en cuenta los valores extremos, mientras que la 2ª trabaja con toda la distribución. Esto, en distribuciones con mucha variación, lleva a problemas de fiabilidad.

4.3 LÍNEAS DE FUTURO.

Ya que se ha demostrado que el método robusto es mejor en muchas condiciones realiastas, sería interesante impulsar una vía diferente en las investigaciones hoy en día, que llevaría a unos resultados más precisos, potentes y fiables.

En el caso de que no podamos comprobar si los datos se ajustan a la curva normal, o que, simplemente si queremos ahorrarnos este esfuerzo, podemos recurrir al empleo de la estadística robusta, que nos llevará a resultados más óptimos aún nuestra distribución sea anormal. De hecho, aun cuando creemos que estamos manejando datos adecuados para solo usar métodos paramétricos, el uso de unos buenos métodos robustos pueden darnos un incremento notable de la exactitud que obtendríamos usando métodos clásicos. Sería necesario estudiar más los métodos robustos para darlos a conocer y potenciar su uso en las investigaciones, con el fin de que no se realicen conclusiones precipitadas e inexactas por el efecto producido por las puntuaciones extremas que pudieran presentarse en los datos a analizar.

En la última edición de Kline (2013) se agrega en cada capítulo información sobre los métodos robustos. Teniendo en cuenta que es un texto editado por la APA, y que representa su apuesta por la reforma de los métodos estadísticos en las ciencias del comportamiento, esto nos da la dimensión de que en el futuro van a ser considerados como alternativa analíticas recomendadas para publicar en revistas científicas de alto impacto.

V. REFERENCIAS BIBLIOGRÁFICAS.

- Carro, J. (1994) Psicoestadística descriptiva. Salamanca: Amarú ediciones.
- Hampel, F., Ronchetti, E., Rousseeuw, P.,Stahel, W. (1986) Robust Statistics. The Approach Based on Influence Funtions. New York: John Wiley.
- Huber, P.J. (1982, 2004). Robust Statistics.New Jersey: Wiley-interscience.
- Kline, R. B. (2013). Beyond significance testing: Statistics reform in the behavioral science (2nd ed.) Washington, DC: American Psichological Association.
- Micceri, T. (1989) The unicorn, the normal curve and other improbable creatures. *PsichologicalBulletin*, vol. 105, 1, 156-166. Recuperado de: http://isites.harvard.edu/fs/docs/icb.topic988008.files/micceri89.pdf
- Norušis, M.J. (1990) The SPSS Guide of Data Analysis. For Release 4. Chicago: SPSS Inc.
- Palmer, A.L (1999). Análisis de Datos, Etapa Exploratoria. Madrid: Pirámide.
- Sociedad Española de Bioquímica Clínica y Patología Molecular (1975) Estadística Robusta. Recuperado de: http://www.seqc.es/dl.asp?175.145.205.255.15.30.27.21.118.133.24.113.255.171.41.12.166.146.68.152.249.7.59.163.205.10.250.118.237.74.68.216.44.202.229.0.69.136.102.106.253.91.165.216.192.188.
- Wilcox, R. (2003) Applying Contemporary Statistical Techniques. San Diego: Academic Press.
- Wilcox, R. (2001) Fundamentals of Modern Statistical Methods. Substantially Improving Power and Accuracy. New York: Springer-Verlag.
- Wilcox, R. (2005) Introduction to Robust Estimation and Hypothesis Testing (2º edition) Burlington: Elsevier Academic Press.
- Zamar, R. (1994) Estimación Robusta. *Estadística Española*, 36 (137), 327-387. Recuperado de: http://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3B+filename%3D843%2F89%2F113_1.pdf&blobkey=urldata&blobtable=MungoBlobs&blobwhere=843%2F89%2F137_1.pdf&ssbinary=true.